A System Outline for Discovery of Proactive and Knowledge Driven Decisions from Virtually Integrated Data Sources using Data Analytics and Big Data Technologies

Proposal submitted for the registration of Ph.D. in Computer Science under the faculty of Science & Technology

Submitted By
Mr. Safwan Abdu Saif Alshaibani
Yemen, Email: Safwansh81@gmail.com

(Through International Student's Centre, SRTMUN)

Research Supervisor
Dr. Parag Bhalchandra
Assistant Professor
SCS, SRTMUN

Research Centre
School of Computational Sciences
S.R.T.M University Nanded, Maharashtra

A System Outline for Discovery of Proactive and Knowledge Driven Decisions from Virtually Integrated Data Sources using Data Analytics and Big Data Technologies

INTRODUCTION:

The present computing scenario is generating mountainous information every day which usually get stored across databases, warehouses and clouds. It's quite interesting to see hidden patterns and proactive knowledge hidden in it. To make this visible, data mining technologies, big data analytics have been evolved provided way through integrated data. Use of extended analytics through above two technologies is an absolutely necessary characteristic for today's businesses processing systems so as to make proactive, knowledge driven decisions and these two technologies help businesses to have future trends and behaviors predicted though voluminous data[1]. There has been the need for professional data processing by the group of science and enterprises. Therefore extreme a number of approaches have been engaged to guarantee that applications can deal with increasing data volumes. With the advancement of Information technologies and applications, a copious amount of data is generated, which attracts both the research community to utilize this information for extracting knowledge and the industry for developing the knowledge-based system. Visualization of data, pattern mining from datasets and analyzing data drift for the different features are three highly used applications of machine learning and data science fields [13]. The rapid development of this sphere is explained by the intermittent growth of information, which leads to the impossibility of applying classical methods and tools for processing available data [15]. Our underlined study plans to work out a system for integration of data analytics and big data technologies with virtual warehouses / data repositories / clouds for discovery of knowledge primarily for business expansions. The term proactive means controlling situations by causing things to happen rather than waiting to respond after things happen. This gives opportunity for constantly moving forward, looking to the future, and making things happen. Being proactive is a way of thinking and acting [3,4]. The Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Danham et al. [7] have proposed a novel, proactive approach to data-mining. This approach involves intervention in the distribution of the input data, with the aim of maximizing an economic utility measure. This intervention requires the consideration of domain-knowledge, which is exogenous to the typical classification task. Them work demonstrates that by taking the proactive approach, it becomes possible to solve business problems that cannot be approach through traditional, passive data-mining methods.

Data mining is not a new concept but a proven technology that has transpired as a key decision-making factor in business. There are numerous use cases and case studies, proving the capabilities of data mining and analysis. Yet, we have witnessed many implementation failures in this field, which can be attributed to technical challenges or capabilities, misplaced business priorities and even clouded business objectives. While some implementations battle through the above challenges, some fail in delivering the right data insights or their usefulness to the

business. At its core, data mining consists of two primary functions, *description*, for interpretation of a large database and *prediction*, which corresponds to finding insights such as patterns or relationships from known values. Before deciding on data mining techniques or tools, it is important to understand the business objectives or the value creation using data analysis. The blend of business understanding with technical capabilities is pivotal in making big data projects successful and valuable to its stakeholders. Business need should be the driving force for implementing data analysis strategy, and technology can be the enabler. Business leaders should know the area of concern they want to tackle using data analytics. Some of the major concerns across business are customer management, revenue growth, bringing operational efficiency and so on. Your business strategy should tackle one major problem at a time, which is then broken down into smaller, specific use cases. Understanding the business value is crucial in identifying and further breaking down business problems. It will form the base of your data strategy which can then be powered by technology to get the desired results. The details of business and technology strategy vary by industries and individual companies, but running through these steps can provide crucial insights on the need for big data analytics. Patrons should ask these questions before embarking on this journey - Do I need this technology? How can it be beneficial? How can I evaluate its effectiveness? These assessments are imperative in defining the need and long-term benefits of data analysis strategy.

In similar way, big data is a term utilized to refer to the increase in the volume of data that are difficult to store, process, and analyze through traditional database technologies. The nature of big data is indistinct and involves considerable processes to identify and translate the data into new insights. The term "big data" is popular in IT and business [2], so we use new mechanism and techniques for processing for these big data to be readable and understandable. In recent years, Big Data analytics has become popular research in computer science and other related data mining areas. It is a very important research area in new trends and also creates enormous vibration in technocrats. It is not only vibrant in computer science but also spread into many other diversified areas. In many industrial applications, the data storage and retrieval after analysis become a primary task in frequent situations. But achieving at correct and accurate data becomes a tedious job at data bases. The Big Data calls for scalable storage index and a distributed approach to retrieve required results [5] and its size is exceeding Terabytes. Because of the variety of data that it encompasses, big data always brings a number of challenges relating to its volume and complexity. A recent survey says that 80% of the data created in the world are unstructured. One challenge is how these unstructured data can be structured. Another challenge is how we can store it. There are the top technologies used to store and analyse Big Data We can categorise them into two (storage and Querying/Analysis) are Apache Hadoop and Microsoft HDInsight [6].

LITERATURE REVIEW:

There exists many contemporary works on data mining, big data analytics and cloud computing, which can be systematically bound into a theory. A work by D.Sawaitul et al. [8] have proposed a novel method for Classification and Prediction of future weather by using Back Propagation Algorithm and the weather parameters like speed of the wind, wind direction, rainfall, temperature and forecasting the climatic conditions are recorded. The artificial neural network back propagation algorithm is used to predict the weather conditions. The Author has tried three models to predict the weather conditions. The first model is used to collect the

weather forecasting techniques. The second model is used to introduce the WSN tool kit for collect the data and then third model is used the Back Propagation Algorithm can be applied on different parameters of weather forecast. The Zenkert et al. [9] have proposed a smart mobility strategy based on the analysis of large datasets which are collected from different sensors, all deployed at various places in a smart city. The data contains various features such as pollution, road traffic, weather, parking, as well as cultural and library event information. The presented smart mobility approach uses the pollution and road traffic information from the data. The analysis and evaluation of the data are done with a Hadoop cluster using MapReduce to concurrently calculate routing possibilities and to find the least polluted way from start to end points based on current conditions indicated by the sensor data. The Tian et al. [10] have proposed a model called Long Short-Term Memory Recurrent Neural Network (LSTM RNN), which takes advantages of the three multiplicative units in the memory block to determine the optimal time lags dynamically. The dataset from Caltrans Performance Measurement System (PeMS) is used for building the model and comparing LSTM RNN with several well-known models, such as random walk (RW), support vector machine (SVM), single layer feed forward neural network (FFNN) and stacked auto-encoder (SAE). The Sciarrone et al.[11] have clarified some aspects concerning different disciplines: machine learning, data mining, educational data mining, teaching analytics and learning analytics with the aim of giving a unique view to many projects that in literature are presented under different points of view. Having a vision not only related to the particular application implemented but also to its location in a cyclic process of learning and teaching improvement is of fundamental importance in order to compare different proposals. The Grady et al. [12] have discussed at a conceptual level the many activities that have been added to our KDDS (Knowledge Discovery in Databases) process model that we use in our customer engagements. The focus has been on the new activities needed for the design and development of big data systems. Some tasks naturally overlap with other disciplines , but the tasks discussed here directly relate to the efficient and accurate delivery of value from data science applications involving big data. The Arafat et al. [13] have proposed such a data visualization tool, named VIM, which is a web-based comprehensive tool for generic data visualization, data preprocessing and mining suitable knowledge with drift analysis of data. Given a dataset, it can envisage the distribution of data with convenient statistical diagrams for different selected features. Moreover, users can employ VIM to generate association rules by selecting multiple features. They have developed VIM using Python Django framework and GraphLab library. The Donghui et al. [14] have constructed a big data analytics framework for forecasting rare customer complaints. First, they built a big data ingestion pipelines on a Hadoop platform: a) Ingest MA plan's customer complaints data from CTM from past 3 years. b) Ingest health plan's call center data for MA members from past 3 years, including both structured data and unstructured text script for the calls. c) Ingest MA members' medical claims, including members' demographics and enrollment history. d) Ingest MA members' pharmacy claims. e) Integrate and unified data from above sources, and enrich the data with additional engineered features into a big wide table, one row per member for analysis and modeling. Second, they designed a unique decision tree based Large Ensemble with Over-Sampling (LEOS) algorithm, which mimics random forest but with extreme oversampling of target class to increase bias, and leverages the parallel computing of Hadoop clusters by generating thousands of fixed size training data sets, and for each such dataset training a decision trees with similar fixed tree structure, and ensemble them. Third, they validated our framework and LEOS learning algorithm with real data. All previous literature reviews which mentioned interested of process to fetch big

data from virtually integrated data and analyze this data, the data is either historical data or knowledge driven, after reduce redundant of data and feature selection to discovery proactive and predictions decisions. These analyzes are applied by algorithms and technologies in order to reach best the results and the speed is characterized by the other studies mentioned above.

PROBLEM STATEMENT:

This research aims at discovery and extracting of proactive and knowledge driven decisions from virtually integrated data sources using data analytics and big data technologies to help organizations/ industries/ companies to focus on the most important information in their databases / warehouses/ clouds, as it helps them have future trends and behaviors predicted. The main aim of the study is to build a system that analyzes data and mining in order to reduce redundant of big data and contribute new techniques for betterment of the field. Moreover, this system will extract and discover proactive decisions from virtually integrated data sources which will raise performance the working within avoidance risks and change work procedures previous. Also the system will give to the business model more opportunities to successful.

The specific terms used in the underlined work includes [15],

1. Virtually Integrated Data Sources : There are main issues and features around VDISs namely : autonomy, data location and flexibility, data presentation,
2. Data Analytics : Research and extract information from various sources to describe the main features of a data set in support of business operations and decision making .
3. Big Data Technologies: in the big data have several technologies to process it for examples: - classification, cluster, association and feature extraction.

Data mining parameters include [15],

1. Association - Looking for patterns where one event is connected to another event.
2. Sequence or path analysis - Looking for patterns where one event leads to another later event
3. Classification - Looking for new patterns
4. Clustering - Finding and visually documenting groups of facts not previously known
5. Forecasting - Discovering patterns in data that can lead to reasonable predictions about the future. This area of data mining is known as predictive analytics.

There are many applications of Data mining in real world As, Hospital, Student Management, Airline Reservation, Forecasting, Biometrics, Mathematics, Geographical, Web Mining, Parallel Processing, Space Organization, Data Integrity, etc. And in which the data mining term is very useful. But how to efficiently implement data mining in the platform of cloud computing. The popular Data mining tools are [15],

1. WEKA – It is java based tool. Free to use. It includes association, classification, clustering, visualization and modeling techniques.
2. Rapid miner – Open source, no-coding required. Java based. It integrates data mining techniques data preprocessing, prediction and visualization

3. R-Programming Tool –It supports data miners for classification, analysis, clustering and time-series data analysis.
4. ORANGE – It is python based tool. It supports machine learning, data analytic, and text analysis.
5. KNIME – GUI supported tool. Popular tool for Data Analysis.
6. SAS– It provides prediction and Visualization of models dimension.

Trendy BIG Data Tools are as below [15],

1. Apache Hadoop - Hadoop is an open source, Java-based. It supports the processing and storage of large data sets in a distributed computing environment. Apache Mahout - Open source software based mostly in Hadoop. It has implementations of machine learning and data mining algorithms: classification, clustering, collaborative filtering and frequent pattern mining.
2. R - Open source. Designed for statistical computing and visualization. Pegasus – Open source. Graph mining system built on top of Map Reduce. It permits discover patterns and anomalies in enormous real-world graphs.
3. SPARK- Spark provides an interface for programming entire clusters with hidden data parallelism and fault-tolerance.

Challenging issues with Big data When managing big data challenges take place in the following areas,

1. Data Capture and Storage.
2. Data Searching and Indexing
3. Data Sharing
4. Data Integration
5. Data Transmission.
6. Data Curation.
7. Data Analysis.
8. Data Visualization.

Cloud computing is a type of Internet- based computing. It delivers common resources and data to PCs and other devices on request.

Private cloud – This model especially operated for an organization. More control and Security

Public cloud – This model made available to the universal public. Free or pay-per - usage. Simple and reasonably priced model.

Hybrid cloud - It is an arrangement of two or more clouds- private, community or public. It can join comparison, achieved and/or devoted services with cloud resources.

Cloud services can be classified into three categories: Infrastructure as a service Platform as a service Software as a service Infrastructure-as-a-service (IaaS) Provide virtualized server and software. It will be hosted inside the Cloud data center. Beneficiaries pay only the agreed on cost. No need to purchase hardware / software and license. Platform as a service (PaaS) has in

addition to the IaaS application layer, which the customer intends to use. Software as a service (SaaS) is offering software to the others remotely as a Web-based service. It allows organizations to access business functionality at a price usually below paying for certified applications and using resources for that application.

SPECIFIC OBJECTIVES OF THE STUDY:

The main objective of this research is to have full understanding for the area of data analytic in big data and contribute new techniques for betterment of the field. Following are the proposed objectives for big data analytics:-

1. To provide an overview of the data analytics and related over virtually integrated data research work done in this area.
2. To explore different techniques already used for adapting data analytics on virtually integrated data sources.
3. To implement a technique for data analytics using big data technologies.
4. To analyzing and evaluating the performance of the implemented models.
5. To characterize new innovative approaches and testing of them, comparison of their efficiencies with existing ones.
6. To give recommendations and conclusion for development of a system for discovery of proactive and knowledge driven decisions from virtually integrated data sources using data analytics and big data technologies and future enhancement.

RESEARCH LIMITATIONS:

Public access to clouds / virtual data sources, structured and unstructured data processing, common formats for datasets. There are some Cloud challenges. The Cloud computing has been moved away from PCs and an enterprise application server to cloud services. The outward show of cloud computing has made a spectacular impact on the industry. Right now industry desires Cloud services for best opportunities to the society. Similarly, there are some Cloud computing issues. We get any new technology, the implementation of cloud computing is not free from issues. Some of the most important issues include, Security, privacy, service quality, interoperability, availability and reliability, portability, performance and bandwidth cost, etc.

RESEARCH METHODOLOGY:

We will take pattern or data from the virtually integrated data source where we will analyze the data in the methods available for analysis and we will use big data technologies. This will result in proactive decisions detection on information knowledge from the data source. We can plan for working on batch data, real time data and various other sources of data**.** Some broad steps includes [15],

1. Once we have figured out the business need, we can indulge in the technical side of data mining steps.
2. Data identification and acquisition is the foremost step for successful implementation. Understanding the business challenges that you are trying to solve helps in determining

the source and types of data to utilize. Data can be in any form - it can be a subset of variables, or can be data samples from a larger database. The key data should directly correlate to the business objective. We will further discuss this in detail while selecting data mining techniques.

3. Data cleansing or cleaning is done on the target data set to improve its effectiveness for fulfilling data mining objectives. This process is primarily to identify inaccurate, incorrect or incomplete data and then replacing, modifying or removing it. This process ensures that your data is complete and error free, thus making it more relevant and effective.
4. Data exploration is at the core of data mining activity. The main objective of this step is to identify the correct data mining techniques or methods and selecting the best suited algorithms for those techniques. Some of the most known data mining techniques include association, classification, regression, segmentation, link analysis, etc. Selecting data mining techniques among the pool is one of the most difficult decisions. However, the selection parameter should consider the objective of business and the available data sets. Mostly, two or more combinations of data mining techniques are used, but it depends on the scale of the project. The last leg of this activity is to perform data mining, i.e., to search for patterns hidden in the data.
5. Data presentation is the final activity which includes interpretation and evaluation of patterns and presenting them to users in a logical and understandable way. The main objective here is to present useful patterns and discard any redundant or irrelevant ones. For any inconsistency, revision of data exploration step is required. The success is attributed to its usefulness in providing insights to decision makers and its relevance to the business objectives.

POSSIBLE OUTCOME :

In this undertaken work, we have identified the challenges, opportunities and issues. We have provided an insight into the possible solutions to these problems even though lot of work is needed to be done in this regard. With the maturity of Cloud computing technologies, Data Mining and Big Data technologies are accelerating in several areas of business, science and engineering to solve data intensive problems. In business models the proactive decisions show mixed results of either weak positive or null relationships. So , we expect to build a system which will help to get good decisions, changing work procedures and result in proactive outcomes such as feedback and social support all these result based on knowlodge driven and hestrical data. In addition, the system will improve the business environment and will work to raise the level of working and increase in the rate of profits and the discovery and avoidance of errors that will occur in the future through data analytics and mining continuously. It is also possible that such proactive decisions have benefits regardless of whether proactive outcomes are achieved. Our chosen work, titled A System Outline for Discovery of Proactive and Knowledge Driven Decisions from Virtually Integrated Data Sources using Data Analytics and Big Data Technologies is a fusion of new paradigms for next generation analytics development, enabling large scale data organization, distribution, knowledge discovery, decision making and penetrating of large volumes rapidly growing diversity forms of data using virtual data sources / Cloud computing as a back end large scale service-oriented computational set-up facility. This paradigm combines huge scale compute, new data exhaustive techniques and precise models to

build data analytics for built-in information extraction. Organizations uphold to store more and more data in cloud environments, which signifies enormous, expensive source of information to mine and clouds offer commercial users scalable resources on request. Cloud computing is emerged as service oriented computing model, to distribute infrastructure, platform and applications as services from the providers to the consumers meeting the Quality of Service parameters by enormous volumes of data at faster scale based on market models. Data mining is a technique that is used to provide data visualization. Big Data demands enormous computing data resources and Clouds present huge scale set-up, hence both these technologies could be combined. Data Mining with big data can be used to paddling through log files, internal strife click streams, transaction analysis, dealing with the social media, to avoid fraud and trying to manage protein series. Cloud computing is a normal fit for Data Mining and Big data analytics. Flexible compute capacity and on-demand provisioning make analytics accessible to more teams within society, while Apache Hadoop has reduced the time to complete analysis. Here design, deployment, operation and Mining big data cloud applications across public and private clouds maintained by Cloud Management. The Cloud architectures includes an arrays of virtual machines that are model for the processing of very huge data sets, to the extent that processing can be segmented into several parallel processes. Mining of data from various data sources is monotonous. And also the big data is stored at cloud environment. Applying data mining techniques with big data and the data is processed using parallel computational methods. Finally, the data is combined, predicted using visualization methods. Beside significant addition to the fundamental concepts and their understanding in the subject of computer science, the topic has importance because many business models can be improved with analytics. We expect that the proposed approach and the possible outcome can be generalized in a number of directions including optimization process.

PROPOSED WORK PLAN

Year-wise plan of work and targets to be achieved

Year 1: Literature Survey and contemporary business models, Theoretical analysis of the existing techniques , their scope and limitations

Year 2: Design and implementation technique to improve the proactive decisions, Preprocessing of the data analysis in the virtually integrated data source and Analytics and Performance evaluation of experimental results in terms of recognition rate and prediction time.

Year 3: Result Analysis and Interpretation of the experimental result, Preparation of final synopsis and thesis and Submission of synopsis and thesis.

REFERENCES:

1. Petre, Ruxandra-Stefania. "Data mining in cloud computing." *Database Systems Journal* 3.3 (2012): 67-71.
2. Hashem, Ibrahim Abaker Targio, et al. "The rise of “big data” on cloud computing: Review and open research issues." *Information systems* 47 (2015): 98-115.
3. Yadav, Surjeet Kumar, Brijesh Bharadwaj, and Saurabh Pal. "Data mining applications: A comparative study for predicting student's performance." *arXiv preprint arXiv:1202.4815*(2012).

4. Williams, Helen M., Sharon K. Parker, and Nick Turner. "Proactively performing teams: The role of work design, transformational leadership, and team composition." *Journal of Occupational and Organizational Psychology* 83.2 (2010): 301-324.
5. Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." *Communications of the ACM*51.1 (2008): 107-113.
6. Chen, CL Philip, and Chun-Yang Zhang. "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data." *Information Sciences* 275 (2014): 314-347.
7. Dahan, Haim, et al. "Proactive data mining using decision trees." *Proactive Data Mining with Decision Trees*. Springer, New York, NY, 2014. 21-33.
8. Sawaitul, Sanjay D., K. P. Wagh, and P. N. Chatur. "Classification and prediction of future weather by using back propagation algorithm-an approach." *International Journal of Emerging Technology and Advanced Engineering* 2.1 (2012): 110-113.
9. Zenkert, Johannes, et al. "Big data analytics in smart mobility: Modeling and analysis of the Aarhus smart city dataset." *2018 IEEE Industrial Cyber-Physical Systems (ICPS)*. IEEE, 2018.
10. Tian, Yongxue, and Li Pan. "Predicting short-term traffic flow by long short-term memory recurrent neural network." *Smart City/SocialCom/SustainCom (SmartCity), 2015 IEEE International Conference on*. IEEE, 2015.
11. Sciarrone, Filippo. "Machine Learning and Learning Analytics: Integrating Data with Learning." *2018 17th International Conference on Information Technology Based Higher Education and Training (ITHET)*. IEEE, 2018.
12. Grady, Nancy W. "KDD meets big data." *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 2016.
13. Arafat, Sk Shariful Islam, et al. "VIM: A Big Data Analytics Tool for Data Visualization and Knowledge Mining." *2017 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*. IEEE, 2017.
14. Wu, Donghui. "A big data analytics framework for forecasting rare customer complaints: A use case of predicting MA members' complaints to CMS." *Big Data (Big Data), 2017 IEEE International Conference on*. IEEE, 2017.
15. R. Kabilan, Dr. N. Jayaveeran, Data Mining with Big Data under Cloud Environment – Opportunities, Issues and Challenges International Journal of Advanced Research in Computer and Communication Engineering, Vol. 6, Issue 2, February 2017

Researcher **Research Supervisor**

Safwan Abdu Saif Alshaibani **Dr. Parag Bhalchandra**